JADH Presentation Application Form

Application Detail

1. The type of presentation (poster, short paper, long paper or panel)

   POSTER

2. A title

   A study on the distribution of coocurrence weight patterns of classical Japanese poetic vocabulary

3. A list of keywords (up to five)

   Japanese literature, classical Japanese poetry, midrange lexical layer, stop word substitution

4. The name, status and affiliation of the presenter (s)

   - Hilofumi Yamamoto, Ph. D. in Linguistics, Professor, Tokyo Institute of Technology
   - Bor Hodošček, Dr. of Engineerings, Associate Professor, Osaka University

5. A contact email address

   - yamagen@ila.titech.ac.jp

6. A postal address

   - Tokyo Institute of Technology: W1-8, 2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8550, Japan

A biography

Hilofumi Yamamoto is Professor at Tokyo Institute of Technology. He earned a Ph. D. in Linguistics at Australian National University and is currently working on the mathematical modeling of vocabulary, linguistic change, and language complexity.

Bor Hodošček is Associate Professor at Osaka University. He earned a Dr. of Engineering at the Tokyo Institute of Technology and is currently working on the quantitative modeling of register in Japanese as well as exploring its role in writing assistance systems. His interests include quantitative linguistics, natural language processing, and educational technology.

A study on the distribution of coocurrence weight patterns of classical Japanese poetic vocabulary

Hilofumi Yamamoto
Tokyo Institute of Technology

Bor Hodošček
Osaka University

2018.5.8

1 Introduction

The present study, ongoing work, focuses on exploring the threshold values which divide words into three groups such as content words, functional words, and inbetween words in classical Japanese text. In terms of content or semantic based analysis we usually take some techniques of data clensing such as eleminations of tags, punctiations, or symbols as a preprocess. Stop word is also a type of tokens to be eliminated since they are comparatively less meaning for content analysis. The list of stop words is commonly used, but has some problems: 1) it is necessary to compile them as a word list in advance; 2) it must be changed depending on the domains of analyses; and 3) it is not centain that which words should be included in a list in terms of the analysis of classical texts.

Our previous study grouped modern Japanese words into low-, mid-, and high-range groups according to their information content given by their term frequency-inverse document frequency (*tf-idf*): low range words corresponded to infrequent and highly topical words, and high range words corresponded to functional words expressing the grammatical relations between words. We, however, do not know which point can automatically classify tokens into low-, mid-, and high-range neatly. It is less conducted on midrange words (Hodošček and Yamamoto 2013).

One of the methods using in Hodošček and Yamamoto (2013) exploits the occurrence not of individual words but of pairwise/co-occurrence patterns such as 'flagrance–flower' relationship revealed that the distribution of co-occurrence weight approximately fits to Gaussian curve in modern Japanese texts. We have not enough examinations to prove if it is applicable to the analysis of classical text as well. However, the distribution fitting to Gaussian curve is one of advantageous features for that purpose. We will attempt to apply the distribution characteristics to the analysis of classical texts in the present study.

2 Methods

We use the Hachidaishū as a material of the present study, which is the eight anthologies compiled by the order of Emperors (ca. 905–1205) and contains about 9,500 poems. We developed the corpora of it and a method of co-occurrence weighting, *cw* (Yamamoto 2006) which calculates the weight of patterns of any two words occurring in a poem sentence similar to the *tf-idf* method (Spärck Jones 1972, Robertson 2004, Manning and Schütze 1999).

$$
\begin{aligned}
w(t,d) &= (1+\log\ \mathit{tf}(t,d)) \cdot \mathit{idf}(t) \\
cw(t_1,t_2,d) &= (1+\log\ \mathit{ctf}(t_1,t_2,d)) \cdot \mathit{cidf}(t_1,t_2) \\
\mathit{cidf}(t_1,t_2) &= \sqrt{\mathit{idf}(t_1)\cdot \mathit{idf}(t_2)} \\
\mathit{idf}(t) &= \log \frac{N}{\mathit{df}(t)}
\end{aligned}
$$

Where, w is weight, t is a token, N is the number of tokens. The function, *idf*, is called "inverse document freuency."(Spärck Jones 1972, Robertson 2004, Manning and Schütze 1999) The function *cw* is called "co-occurrence weight," which allows us to examine the patterns of poetic word constructions through mathematical modeling.

As in Figure 1, there is a concept (Losee 2001: 1019) of terms located in each layer being effective query terms. Luhn (1968) cuts the top and bottom words of the frequency and uses midrange vocabulary for development of the automatic outline generation system (Figure 1). Nagao (1983: 28) also mentioned midrange vocabulary effective in generating automatic abstract. Nagao (1983)'s viewpoint is slightly different with Luhn (1968) in that it allocates the distribution of word lengths around the Gaussian curve. The positions both upper-cutoff and lower-cutoff are, however, assumed to be empirical; it is not discussed where to cut them off.

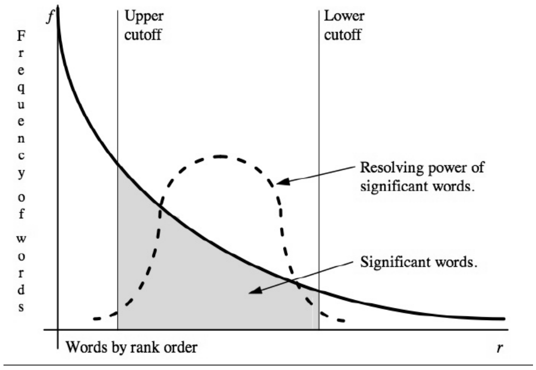

Figure 1: Hyperbolic curve relating occurrence frequency with rank order; adapted from (Luhn 1968: 120)

Table 1: Upper cutoff patterns of *ame* (sakura): *cw* = co-occurrence weight; *z* = z-value. word annotations: ari(be), ba(cond.), ha(topic.), hana(flower), hito(human), keri(past.), ki(past.), koso(emphatic.), miru(see), mo (also), nasi(no exist), nu(neg.), o(obj.), omou(think), ramu(aux.will), su(do), te(p.), to(and), ware(we), zo(emphatic.), zu(neg.).

	cw	*z*	pattern		*cw*	*z*	pattern		*cw*	*z*	pattern
1	0.62	-0.91	mo–keri	11	0.59	-0.96	nasi–ha	21	0.52	-1.05	nu–o
2	0.62	-0.92	hana–o	12	0.57	-0.98	o–ramu	22	0.52	-1.05	o–zo
3	0.62	-0.92	o–koso	13	0.57	-0.98	mo–ramu	23	0.52	-1.05	miru–o
4	0.60	-0.94	zu–keri	14	0.57	-0.98	ha–ki	24	0.48	-1.09	ba–mo
5	0.60	-0.94	su–ha	15	0.56	-1.00	zu–mo	25	0.48	-1.09	o–keri
6	0.60	-0.94	to–ba	16	0.56	-1.00	o–te	26	0.43	-1.16	zu–ha
7	0.59	-0.96	ari–ha	17	0.55	-1.01	hito–mo	27	0.43	-1.16	to–o
8	0.59	-0.96	ari–mo	18	0.54	-1.02	zu–te	28	0.43	-1.16	te–ha
9	0.59	-0.96	ware–mo	19	0.52	-1.05	zo–ha	29	0.34	-1.27	o–ha
10	0.59	-0.96	nasi–o	20	0.52	-1.05	omou–o	30	0.34	-1.27	o–mo

3 Results

The distribution of *cw* values is taken from the network model of both *ume* (plum) and *sakura* (cherry) and their curves belong to Gaussian curve as well as in classical texts (Figure 2). Therefore we will attempt to divid this shape into three layers by inflection points.

The co-occurrence patterns of *sakura* (cherry) under -0.9 (near -1) *cw* value are ajacent patterns comprising function words, and over 1 *cw* value are those of the patterns with content words as we expected (Table 1 and 2). As upper-cutoff, we used under -0.9 (near -1) σ value of *cw*, which could extract patterns of functional tokens: almost all patterns included functional words, while as lower-cutoff, we used over 1 σ values, which could extract patterns of content tokens: almost all patterns included content words. Both under -1 and over 1 σ are regarded as inflection points which have mathematically

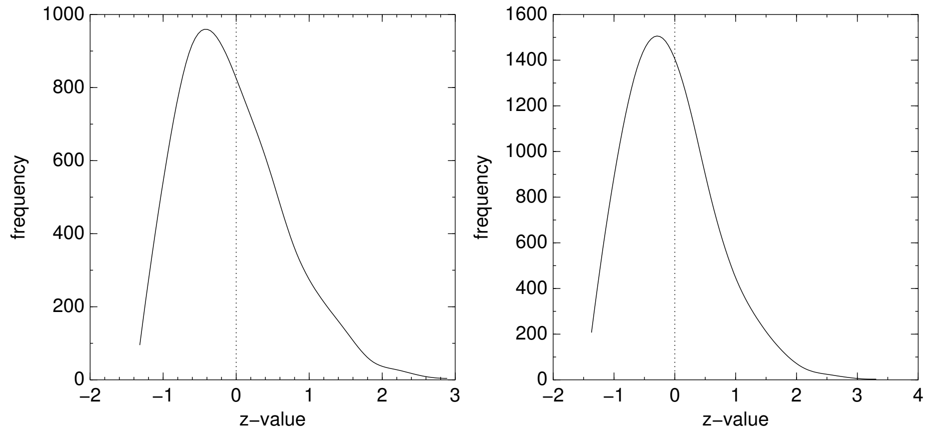

Figure 2: The distribution of cw values *ume* (plum; left) and *sakura* (cherry; right) in Hachidaishū; The statistics of *ume* (plum): N=7016, min=-1.370, mean=0.138, max=3.700, SD=0.740, SE=0.009, CV=534.012%, Reliable interval low - upper = 0.116 - 0.161 (95%), skew=0.737, kurtosis=3.567, and that of *sakura* (cherry): N=4734, min=-1.320, mean=0.132, max=3.240, SD=0.716, SE=0.010, CV=544.116%, Reliable interval low - upper = 0.104 - 0.159 (95%), skew=0.740, kurtosis=3.345 indicate the both approximately fitting to Gaussian curve.

interesting property.

4 Discussion

Inflection points is defined as the points of the curve where the curvature changes its sign while a tangent exists.(Bronshtein et al. 2004: 231) We consider the threshold values which part upper-cutoff, midrange, and lower-cutoff not as coincident but as evidential points. It is, however, necessary to conduct further experiments and continue to discuss its mathematical traits behind the distributions of co-occurrence weight.

In terms of removing low range (upper cutoff) and extracting high range (lower cutoff) from poetic texts, we found that we do not need to use any filters to eliminate terms since *cw* values returned semantically co-occurrence patterns. Apart from low range and high range, it is, however, still unknown the characteristics of midrange lexical layer.

Table 2: Lower cutoff patterns of *ame* (sakura) in Kokinshū: 30 out of 164 patterns extracted; *cw* = co-occurrence weight; *z* = z-value. word annotations: ba(cond.), bakari(only), besi(should be), chiru(fall), fukakusa(deepgreen), hana(flower), isa(already), kakusu(hide), katu(win), koku(pull), komoru(go deep inside), magiru(mix), makasu(entrust), maku(wind up), manimani(as it is), masi(as), mazu(mix), me(eye), minami(south), miyako(city), mono(thing), nagara(even if), sakura(cherry), si(emphasic.), sumi(black ink), tatu(start,stand), tazumu(being around), tu(past.), uturou(change), watasu(give), yamakaze(mountain wind), yamu(stop), yanagi(willow), yononaka(world).

	cw	*z*	pattern		*cw*	*z*	pattern
1	3.86	3.18	yamu–manimani	106	2.38	1.31	si–fukakusa
2	3.75	3.04	minami–magiru	107	2.38	1.31	sakura–hana
3	3.67	2.93	minami–maku	108	2.38	1.31	sakura–isa
4	3.61	2.86	maku–magiru	109	2.38	1.31	sakura–ba
5	3.42	2.62	yanagi–koku	110	2.38	1.30	sakura–me
6	3.38	2.57	yamu–makasu	—			
7	3.38	2.56	mazu–koku	155	2.17	1.04	chiru–katu
8	3.27	2.43	yanagi–mazu	156	2.17	1.04	bakari–sumi
9	3.26	2.42	sakura–yamu	157	2.16	1.03	maku–besi
10	3.25	2.40	minami–yamakaze	158	2.16	1.03	tatu–maku
–				159	2.16	1.03	tatu–tazumu
101	2.40	1.33	uturou–komoru	160	2.16	1.03	tazumu–tu
102	2.40	1.33	sakura–watasu	161	2.16	1.03	miyako–sakura
103	2.40	1.33	katu–nagara	162	2.16	1.02	kakusu–si
104	2.39	1.32	sakura–masi	163	2.14	1.00	yononaka–sakura
105	2.39	1.31	sakura–makasu	164	2.14	1.00	mono–sakura

5 Conclusion

To classify co-occurrence patterns into three divisions, we used one of the distribution characteristics of co-occurrence weight, and we could divide them into three layers of co-occurrence patterns: high, mid, and low range patterns. We found that 1) the distribution of classical texts fits to Gaussian curve as well as of modern texts; 2) *cw* value can separate patterns into three layers (low-, mid-, and high range) by inflection points (-1σ and 1σ); 3) one of the three layers, high range could be extracted without the list of stop words; 4) midrange lexical layer might include mathematical traits, which has not been unveiled yet in the present study.

References

Bronshtein, I.N., K. A. Semendyayev, G. Musiol, and H. Muehlig (2004) *Handbook of Mathematics*: Springer-Verlag, 4th edition.

Hodošček, Bor and Hilofumi Yamamoto (2013) "Analysis and Application of Midrange Terms of Modern Japanese", in *Computer and Humanities 2013 Symposium Proceedings*, No. 4, pp. 21–26.

Losee, Robert M. (2001) “Term dependence: A basis for Luhn and Zipf models”, *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 12, pp. 1019–1025.

Luhn, Hans Peter (1968) *HP Luhn: Pioneer of Information Science: Selected Works*: Spartan Books.

Manning, Christopher D. and Hinrich Schütze (1999) *Foundation of statistical natural language processing*, Cambridge, Massachusetts: The MIT press.

Nagao, Makoto (1983) *Gengo kogaku (Lanuage Engineering)*, Jinkochino sirizu 2 (Series of Artificial Intelligence): Shokodo.

Robertson, Stephen (2004) “Understanding inverse document frequency: on theoretical arguments for IDF”, *Journal of Documentation*, Vol. 60, pp. 503–520.

Spärck Jones, Karen (1972) “A Statistical Interpretation of Term Specificity and Its Application in Retrieval”, *Journal of Documentation*, Vol. 28, pp. 11–21.

Yamamoto, Hilofumi (2006) “Konpyūta niyoru utamakura no bunseki / A Computer Analysis of Place Names in Classical Japanese Poetry”, in *Atti del Terzo Convegno di Linguistica e Didattica Della Lingua Giapponese, Roma 2005*: Associazione Italiana Didattica Lingua Giapponese (AIDLG), pp. 373–382.